

Weekly Report

April 21, 2019

1 Work

1. 本周同时在进行unpair setting下的图片增强，和基于空间位移的Adversarial Attack，目前还在测试思路的可能性。
2. 本周完成了IJCAI的rebuttal。
3. 工作时长：工作日每天11个小时，周末共10个小时，共65个小时。

1.1 工作进度

Table 1: 工作进度

项目	进度	截止时间
DRGraph	需要对程序做一些修改	2019.4.30
unpair 低光照图片增强	目前初步的实验效果不佳	
NIPS	Adversarial Attack	2019.5.23

2 Paper Reading

2.1 ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD (Workshop track - ICLR 2017)

选定一个目标label，计算与目标想对应的梯度，然后加到输入图片中。 $X = X - \text{sign}(d_X \text{loss}(f(X), \text{label}))$. 等价于decrease the cost of the target class。

2.2 Intriguing properties of neural networks (L-BFGS)

Minimize $c|r| + \text{loss}f(x + r, l)$ 使得图片x加上对抗扰动r之后得到的label是给定的错误标签l。



Figure 1: #2

2.3 One Pixel Attack for Fooling Deep Neural Networks

只修改一个像素，是的图片的分类结果错误。 $|v|_0 \leq 1$

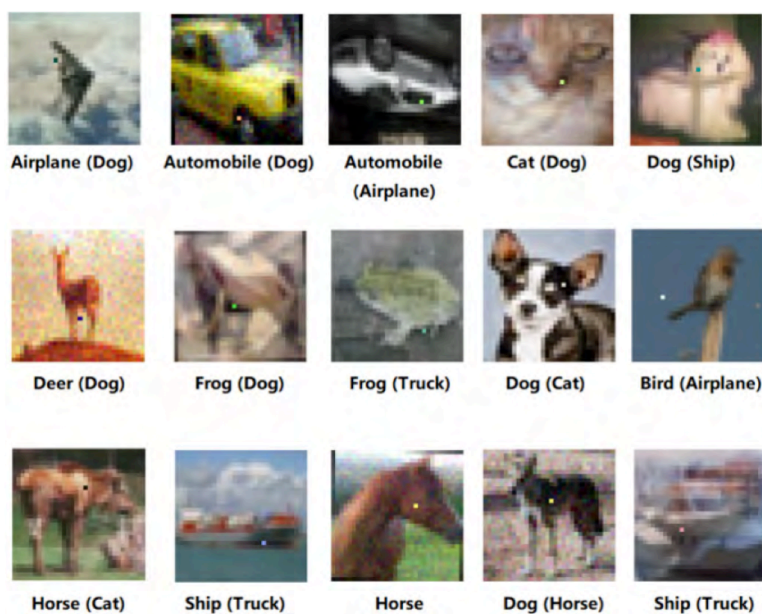


Figure 2: #3

2.4 DeepFool: a simple and accurate method to fool deep neural networks

Deepfool把分类器近似看做是一个分类平面，那么对抗扰动就是使得 $x + \epsilon$ 可以最快得到达某一个分类平面上，也就是说我们只要计算 $f(x)$ 到分类平面 $f(x)=0$ 的距离即可。

Algorithm 1 DeepFool for binary classifiers

```
1: input: Image  $\mathbf{x}$ , classifier  $f$ .  
2: output: Perturbation  $\hat{\mathbf{r}}$ .  
3: Initialize  $\mathbf{x}_0 \leftarrow \mathbf{x}$ ,  $i \leftarrow 0$ .  
4: while  $\text{sign}(f(\mathbf{x}_i)) = \text{sign}(f(\mathbf{x}_0))$  do  
5:    $\mathbf{r}_i \leftarrow -\frac{f(\mathbf{x}_i)}{\|\nabla f(\mathbf{x}_i)\|_2^2} \nabla f(\mathbf{x}_i)$ ,  
6:    $\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \mathbf{r}_i$ ,  
7:    $i \leftarrow i + 1$ .  
8: end while  
9: return  $\hat{\mathbf{r}} = \sum_i \mathbf{r}_i$ .
```

Figure 3: #4